"VOICE OPERATED WHEELCHAIR FOR PHYSICALLY DISABLED"

By

Kuldeep Naruka (111047)

Saurabh Pal (111050)

Vivek Kumar (111039)

Under the supervision of

Dr. Neeru Sharma

May-2015

Dissertation submitted in partial fulfilment

of the requirement for the Degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT, SOLAN- 173234, INDIA

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

(Established under the Act 14 of Legislative Assembly of Himachal Pradesh)

Waknaghat, P.O. DomeharBani. Teh. Kandaghat, Distt. Solan- 173234(H.P.)

Phone: 01792-245367, 245368, 245369

Fax- 01792-245362

CERTIFICATE

This is to certify that the project entitled "**VOICE OPERATED WHEEL CHAIR FOR PHYSICALLY DISABLED**" submitted by "**Mr Kuldeep Singh Naruka, Mr Saurabh Pal and Mr Vivek Kumar"** to the Department of Electronics Communication Engineering, Jaypee University of Information Technology, Waknaghat, in the partial fulfillment of the degree of Bachelor of Technology in Electronics Communication Engineering, during the year 2011-2015 is an authentic record of the work carried out under my supervision and guidance.

Date:

Dr. Neeru Sharma

(Assistant Professor)

Department of Electronics and Communication Engineering

Jaypee University of Information Technology (JUIT)

Waknaghat, Solan- 173234, (H.P.) India

(Supervisor)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

(Established under the Act 14 of Legislative Assembly of Himachal Pradesh)

Waknaghat, P.O. DomeharBani. Teh. Kandaghat, Distt. Solan- 173234(H.P.)

Phone: 01792-245367, 245368, 245369

Fax- 01792-245362

DECLARATION

We hereby declare that the work reported in the B. Tech thesis entitled " **VOICE OPERATED WHEEL CHAIR FOR PHYSICALLY DISABLED** " submitted by **"Mr Kuldeep Singh Naruka, Mr Saurabh Pal and Mr Vivek Kumar"** at Jaypee University Of Information Technology, Waknaghat is an authentic record of our work carried out under the supervision of **Dr. Neeru Sharma**. This work has not been submitted partially or wholly to any other university or institution for the award of this or any other degree or diploma.

Kuldeep Naruka (111047)

Saurabh Pal (111050)

Vivek Kumar (111039)

Department of Electronics and Communication Engineering

Jaypee University of Information Technology (JUIT)

Waknaghat, Solan- 173234, India

ACKNOWLEDGEMENT

After the competitions of our thesis work, we feel to convey our indebtedness to all those who helped us to reach our goal. We take this opportunity to express our profound gratitude and deep regards to our guide **Dr. Neeru Sharma** for her exemplary guidance, monitoring and constant encouragement throughout the course of this Project. The blessing, help and guidance given by her time to time shall carry us a long way in the journey of life on which we are about to embark. We are obliged to our faculty members of JUIT, for the valuable information provided by them in their respective fields. We are grateful for their cooperation during the period of our Project.

Contents

List of Figures

ABSTRACT

A conventional wheel chair is manually operated and always requires a person to carry the disabled, considering this drawback we brought the idea of a wheel chair which is controlled by using speech signal. This solves the problem of relying on others for ones movement. In the voice operated wheel chair the input speech signal is given through micro phone. Speech extractor is used to convert the given speech signal to word signal. The extracted word signal is recognized by using the speech recognizer. The word signal generates the command. According to the generated command the various operations were performed in the mobile robot like move forward, backward, left and right, clockwise rotate, anticlockwise rotate, open, close, up, down, and stop. The methods used in this research are Linear Predictive Coding (LPC) and Hidden Markow Model (HMM).LPC is used to extract word data from speech signal. HMM is used to recognize the word pattern data, which are extracted from a speech signal .Sampling rate of the speech signal is 8kHz.The use of intelligent wheelchair encourages the view of the machine as a partner rather than as a tool. These people, suffering from disorientation, amnesia, or cognitive deficits, are dependent upon others to push them, so often feel powerless and out of control. Intelligent wheelchair has the potential to provide these people with effective ways to alleviate the impact of their limitations.

Chapter 1

Introduction

The fundamental purpose of speech is communication, the transmission of messages. According to Shannon's information theory, a message represented as a sequence of discrete symbols can he quantified by its information content in bits, and the rate of transmission of information is measured in bits/second (bps). In speech production, as well as in many human-engineered electronic communication systems, the information to be transmitted is encoded in the form of a continuously Varying (analog)waveform that can be transmitted, recorded, manipulated, and ultimately decoded by a human listener.

In the case of speech, the fundamental analog form of the message is an acoustic waveform, which we call the speech signal. Speech signals can be converted to an electrical waveform by a microphone, further manipulated by both analog and digital signal processing , and then converted back to acoustic form by a loudspeaker, a telephone handset or headphone, as desired. This form of speech processing is, of course, the basis for Bells telephone invention as well as today's multitude of devices for recording, transmitting, and manipulating speech and audio signals.

Although Bell made his invention without knowing the fundamentals of information theory, these ideas have assumed great importance in the design of sophisticated modern communications systems.

Therefore, even though our main focus will be mostly on the speech waveform and its representation in the form of parametric models, it is nevertheless useful to begin with a discussion of how information is encoded in the speech waveform.

As relevant background to the field of speech recognition, this chapter intends to discuss how the speech signal is produced and perceived by human beings. This is an essential subject that has to be considered before one can pursue and decide which approach to use for speech recognition.

1.1 The Speech Signal

Human communication is to be seen as a comprehensive diagram of the process from speech production to speech perception between the talker and listener as shown in Figure 1.1.

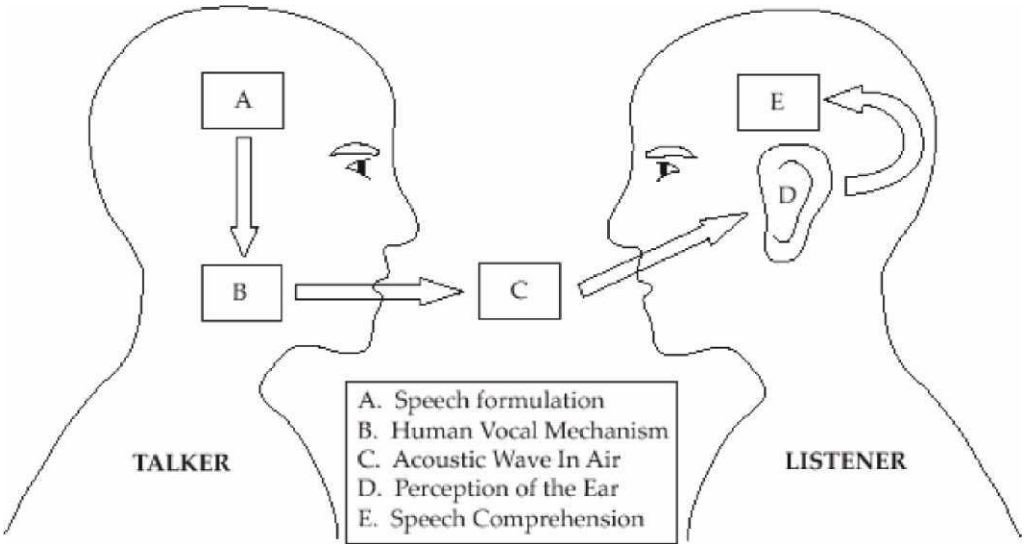

Figure 1.1: Schematic Diagram of the Speech Production/Perception Process

Five different elements:

A. Speech formulation

B. Human vocal mechanism

C. Acoustic air

D. Perception of the ear

E. Speech comprehension.

The first element (A. Speech formulation) is associated with the formulation of the speech signal in the talker's mind. This formulation is used by the human vocal mechanism (B. Human vocal mechanism) to produce the actual speech waveform.

The waveform is transferred via the air (C. Acoustic air) to the listener. During this transfer the acoustic wave can be affected by external sources, for example noise, resulting in a more complex waveform. When the wave reaches the listener's hearing system (the ears) the listener percepts the

waveform (D. Perception of the ear) and the listener's mind (E. Speech comprehension) starts processing this waveform to comprehend its content so the listener understands what the talker is trying to tell him. One issue with speech recognition [1] is to "simulate" how the listener process the speech produced by the talker. There are several actions taking place in the listeners head and hearing system during the process of speech signals. The perception process can be seen as the inverse of the speech production process.

The basic theoretical unit for describing how to bring linguistic meaning to the formed speech, in the mind, is called phonemes. Phonemes can be grouped based on the properties of either the time waveform or frequency characteristics and classified in different sounds produced by the human vocal tract. Speech is:

- Time-varying signal,
- Well-structured communication process,
- Depends on known physical movements,
- Composed of known, distinct units (phonemes),
- Is different for every speaker,
- May be fast, slow, or varying in speed,
- May have high pitch, low pitch, or be whispered,
- Has widely-varying types of environmental noise,
- May not have distinct boundaries between units (phonemes),
- Has an unlimited number of words.

1.2 Speech Production

To be able to understand how the production of speech is performed one need to know how the human's vocal mechanism is constructed as shown in Figure 1.2. The most important parts of the human vocal mechanism are the vocal tract together with nasal cavity, which begins at the velum. The velum is a trapdoor-like mechanism that is used to formulate nasal sounds when needed. When the velum is lowered, the nasal cavity is coupled together with the vocal tract to formulate the desired speech signal. The cross-sectional area of the vocal tract is limited by the tongue, lips, jaw and velum and varies from 0-20 cm^2.

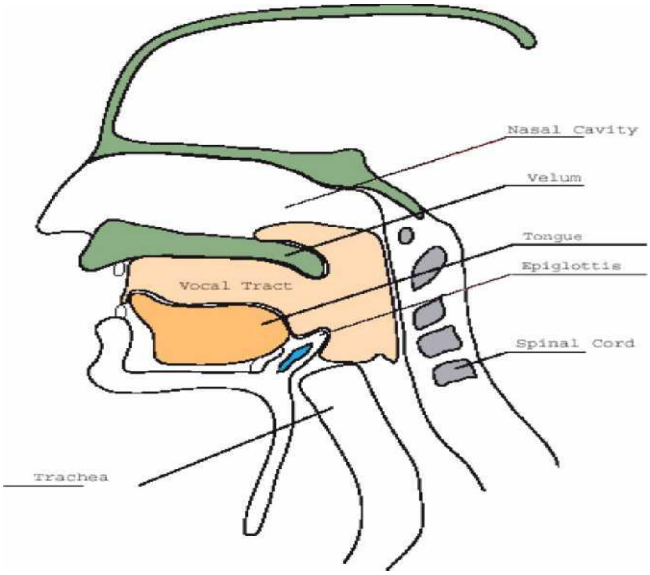

Figure 1.2: Human Vocal Mechanism

1.3 Properties of Human Voice

One of the most important parameter of sound is its frequency. The sounds are discriminated from each other by the help of their frequencies. When the frequency of a sound increases, the sound gets high-pitched and irritating. When the frequency of a sound decreases, the sound gets deepen. Sound waves are the waves that occur from vibration of the materials. The highest value of the frequency that a human can produce is about 10 kHz. And the lowest value is about 70 Hz. These are the maximum and minimum values. This frequency interval changes for every person. And the magnitude of a sound is expressed in decibel (dB). A normal human speech has a frequency interval of 100 Hz – 3200 Hz and its magnitude is in the range of 30 dB - 90 dB. A human ear can perceive sounds in the frequency range between 16 Hz and 20 kHz and a frequency change of 0.5 % is the sensitivity of a human ear.

Speaker Characteristics,

- Due to the differences in vocal tract length, male, female, and children's speech are different.
- Regional accents are the differences in resonant frequencies, durations, and pitch.
- Individuals have resonant frequency patterns and duration patterns that are unique (allowing us to identify speaker).
- Training on data from one type of speaker automatically "learns" that group or person's characteristics, makes recognition of other speaker types much worse.

Chapter 2

Automatic Speech Recognition System (ASR)

2.1 Introduction

Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation whereby speech processing can be seen as the interaction of digital signal processing and natural language processing. Natural language processing is a subfield of artificial intelligence and linguistics. It studies the problems of automated generation and understanding of natural human languages. Natural language generation systems convert information from computer databases into normal-sounding human language, and natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.

2.1.1 Speech coding

It is the compression of speech (into a code) for transmission with speech codec that use audio signal processing and speech processing techniques. The techniques used are similar to that in audio data compression and audio coding where knowledge in psychoacoustics is used to transmit only data that is relevant to the human auditory system. For example, in narrow band speech coding, only information in the frequency band of 400 Hz to 3500 Hz is transmitted but the reconstructed signal is still adequate for intelligibility.

However, speech coding differs from audio coding in that there is a lot more statistical information available about the properties of speech. In addition, some auditory information which is relevant in audio coding can be unnecessary in the speech coding context. In speech coding, the most important criterion is preservation of intelligibility and "pleasantness" of speech, with a constrained amount of transmitted data.

It should be emphasized that the intelligibility of speech includes, besides the actual literal content, also speaker identity, emotions, intonation, timbre etc. that are all important for perfect intelligibility. The more abstract concept of pleasantness of degraded speech is a different property

than intelligibility, since it is possible that degraded speech is completely intelligible, but subjectively annoying to the listener.

2.1.2 Speech synthesis

Speech synthesis is the artificial production of human speech. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. Synthesized speech can also be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diaphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

The quality of a speech synthesizer is judged by its similarity to the human voice, and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1980s.

2.1.3 Voice analysis

Voice problems that require voice analysis most commonly originate from the vocal cords since it is the sound source and is thus most actively subject to tiring. However, analysis of the vocal cords is physically difficult. The location of the vocal cords effectively prohibits direct measurement of movement. Imaging methods such as x-rays or ultrasounds do not work because the vocal cords are surrounded by cartilage which distorts image quality. Movements in the vocal cords are rapid, fundamental frequencies are usually between 80 and 300 Hz, thus preventing usage of ordinary video. High-speed videos provide an option but in order to see the vocal cords the camera has to be positioned in the throat which makes speaking rather difficult. Most important indirect methods are inverse filtering of sound recordings and electro-glotto-graphs (EGG). In inverse filtering methods, the speech sound is recorded outside the mouth and then filtered by a mathematical method to remove the effects of the vocal tract. This method produces an estimate of the waveform of the pressure pulse which again inversely indicates the movements of the vocal cords. The other kind of inverse indication is the electro-glotto-graphs, which operates with electrodes attached to the subject's throat close to the vocal cords. Changes in conductivity of the throat indicate inversely how large a portion of the vocal cords are touching each other. It thus yields one-dimensional

information of the contact area. Neither inverse filtering nor EGG is thus sufficient to completely describe the glottal movement and provide only indirect evidence of that movement.

2.1.4 Speech recognition

Speech recognition is the process by which a computer (or other type of machine) identifies spoken words. Basically, it means talking to your computer, and having it correctly recognize what you are saying. This is the key to any speech related application.

As shall be explained later, there are a number ways to do this but the basic principle is to somehow extract certain key features from the uttered speech and then treat those features as the key to recognizing the word when it is uttered again.

2.2 Speech Recognition Basics

2.2.1 Utterance

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences.

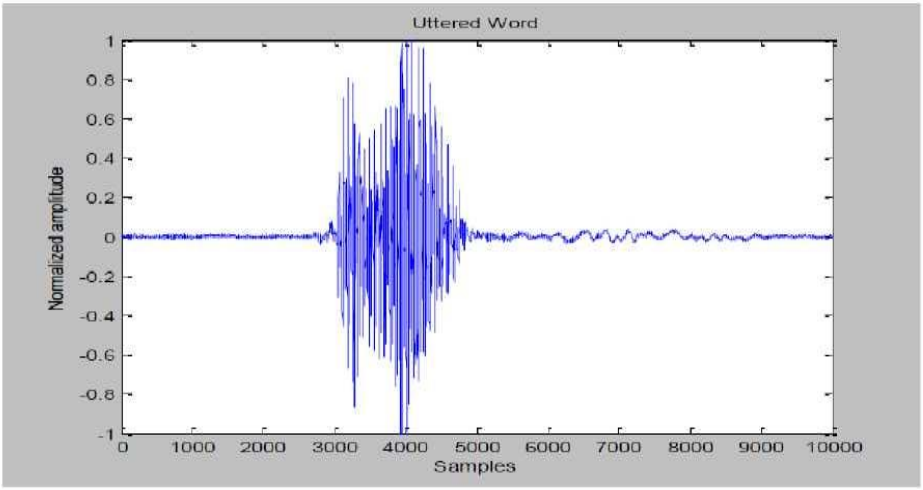

Figure 2.1: Utterance of "HELLO"

2.2.2 Speaker Dependence

Speaker dependent systems are designed around a specific speaker. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy.

2.2.3 Vocabularies

Vocabularies (or dictionaries) are lists of words or utterances that can be recognized by the SR system. Generally, smaller vocabularies are easier for a computer to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. They can be as long as a sentence or two. Smaller vocabularies can have as few as 1 or 2 recognized utterances (e.g." Wake Up"), while very large vocabularies can have a hundred thousand or more!

2.2.4 Accuracy

The ability of a recognizer can be examined by measuring its accuracy - or how well it recognizes utterances. This includes not only correctly identifying an utterance but also identifying if the spoken utterance is not in its vocabulary. Good ASR systems have an accuracy of 98% or more! The acceptable accuracy of a system really depends on the application.

2.2.5 Training

Some speech recognizers have the ability to adapt to a speaker. When the system has this ability, it may allow training to take place. An ASR system is trained by having the speaker repeat standard or common phrases and adjusting its comparison algorithms to match that particular speaker. Training a recognizer usually improves its accuracy.

Training can also be used by speakers that have difficulty speaking, or pronouncing certain words. As long as the speaker can consistently repeat an utterance, ASR systems with training should be able to adapt.

2.3 Classification of ASR System

A speech recognition system can operate in many different conditions such as speaker dependent/independent, isolated/continuous speech recognition, for small/large vocabulary. Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are based on the fact that one of the difficulties of ASR is the ability to determine when a speaker starts and finishes an utterance. Most packages can fit into more than one class, depending on which mode they're using.

2.3.1 Isolated Words

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on BOTH sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. Often, these systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

2.3.2 Connected Words

Connect word systems (or more correctly 'connected utterances') are similar to Isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

2.3.3 Continuous Speech

Continuous recognition is the next step. Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation.

2.3.4 Spontaneous Speech

There appears to be a variety of definitions for what spontaneous speech [5] actually is. At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

2.3.5 Speaker Dependence

ASR engines can be classified as speaker dependent and speaker independent. Speaker Dependent systems are trained with one speaker and recognition is done only for that speaker. Speaker Independent systems are trained with one set of speakers. This is obviously much more complex than speaker dependent recognition. A problem of intermediate complexity would be to train with a group of speakers and recognize speech of a speaker within that group. We could call this speaker group dependent recognition.

2.4 Why is Automatic Speaker Recognition hard?

There are a few problems in speech recognition that haven't yet been discovered. However there are a number of problems that have been identified over the past few decades most of which still remain unsolved. Some of the main problems in ASR are:

2.4.1 Determining word boundaries

Speech is usually continuous in nature and word boundaries are not clearly defined. One of the common errors in continuous speech recognition is the missing out of a minuscule gap between words. This happens when the speaker is speaking at a high speed.

2.4.2 Varying Accents

People from different parts of the world pronounce words differently. This leads to errors in ASR. However this is one problem that is not restricted to ASR but which plagues human listeners too.

2.4.3 Large vocabularies

When the number of words in the database is large, similar sounding words tend to cause a high amount of error i.e. there is a good probability that one word is recognized as the other.

2.4.4 Changing Room Acoustics

Noise is a major factor in ASR. In fact it is in noisy conditions or in changing room acoustic that the limitations of present day ASR engines become prominent.

2.4.5 Temporal Variance

Different speakers speak at different speeds. Present day ASR engines just cannot adapt to that.

2.5 Speech Analyzer

Speech analysis, also referred to as front-end analysis or feature extraction, is the first step in an automatic speech recognition system. This process aims to extract acoustic features from the speech waveform. The output of front-end analysis is a compact, efficient set of parameters that represent the acoustic properties observed from input speech signals, for subsequent utilization by acoustic modeling.

There are three major types of front-end processing techniques, namely linear predictive coding (LPC), Time Domain Analysis(TDA), mel-frequency cepstral coefficients (MFCC) [2], and perceptual linear prediction (PLP), where the latter two are most commonly used in state-of-the-art ASR systems.

2.5.1 Linear predictive coding

LPC starts with the assumption that a speech signal is produced by a buzzer at the end of a tube (voiced sounds), with occasional added hissing and popping sounds. Although apparently crude, this model is actually a close approximation to the reality of speech production. The glottis (the space between the vocal cords) produces the buzz, which is characterized by its intensity (loudness) and frequency (pitch). The vocal tract (the throat and mouth) forms the tube, which is characterized by its resonances, which are called formants. Hisses and pops are generated by the action of the tongue, lips and throat during sibilants and plosives. LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz.

The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue. The numbers which describe the intensity and frequency of the buzz, the formants, and the residue signal, can be stored or transmitted somewhere else. LPC synthesizes the speech signal by reversing the process: use the buzz parameters and the residue to create a source signal, use the formants to create a filter (which represents the tube), and run the source through the filter, resulting in speech. Because speech signals vary with time, this process is done on short chunks of the speech signal, which are called frames; generally 30 to 50 frames per second give intelligible speech with good compression.

2.5.2 Mel Frequency Cepstrum Coefficients

These are derived from a type of cepstral representation of the audio clip (a "spectrum-of-a-spectrum"). The difference between the cepstrum and the Mel-frequency cepstrum [4] is that in the MFC, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly-spaced frequency bands obtained directly from the FFT or DCT. This can allow for better processing of data, for example, in audio compression.

However, unlike the sonogram, MFCCs lack an outer ear model and, hence, cannot represent perceived loudness accurately. MFCCs are commonly derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal
2. Map the log amplitudes of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take the Discrete Cosine Transform of the list of Mel log-amplitudes, as if it were a signal.
4. The MFCCs are the amplitudes of the resulting spectrum.

2.5.3 Perceptual Linear Prediction

Perceptual linear prediction, similar to LPC analysis, is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction (PLP) modifies the short-term spectrum of the speech by several psychophysically based transformations. This technique uses three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum:

(1) The critical-band spectral resolution,

(2) The equal-loudness curve, and

(3) The intensity-loudness power law.

The auditory spectrum is then approximated by an autoregressive all-pole model. In comparison with conventional linear predictive (LP) analysis, PLP analysis is more consistent with human hearing.

2.6 Speech Classifier

The problem of ASR belongs to a much broader topic in scientific and engineering so called *pattern recognition.* The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called *patterns* and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques

described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as *feature matching.*

The state-of-the-art in feature matching techniques used in speaker recognition includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ).

2.6.1 Dynamic Time Warping

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another they were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics -indeed, any data which can be turned into a linear representation can be analyzed with DTW.

A well known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models.

2.6.2 Hidden Markov Model

The basic principle here is to characterize words into probabilistic models wherein the various phonemes which contribute to the word represent the states of the HMM [6] while the transition probabilities would be the probability of the next phoneme being uttered (ideally 1.0). Models for the words which are part of the vocabulary are created in the training phase. Now, in the recognition phase when the user utters a word it is split up into phonemes as done before and it’s HMM is created. After the utterance of a particular phoneme, the most probable phoneme to follow is found from the models which had been created by comparing it with the newly formed model. This chain from one phoneme to another continues and finally at some point we have the most probable word out of the stored words which the user would have uttered and thus recognition is brought about in a finite vocabulary system. Such a probabilistic system would be more efficient than just cepstral analysis as these is some amount of flexibility in terms of how the words are uttered by the users.

2.6.3 Vector Quantization (VQ)

VQ [3] is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a centroid. The collection of all codeword's is called a codebook.

Figure 2.2 shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result codewords (centroids) are shown in Figure by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

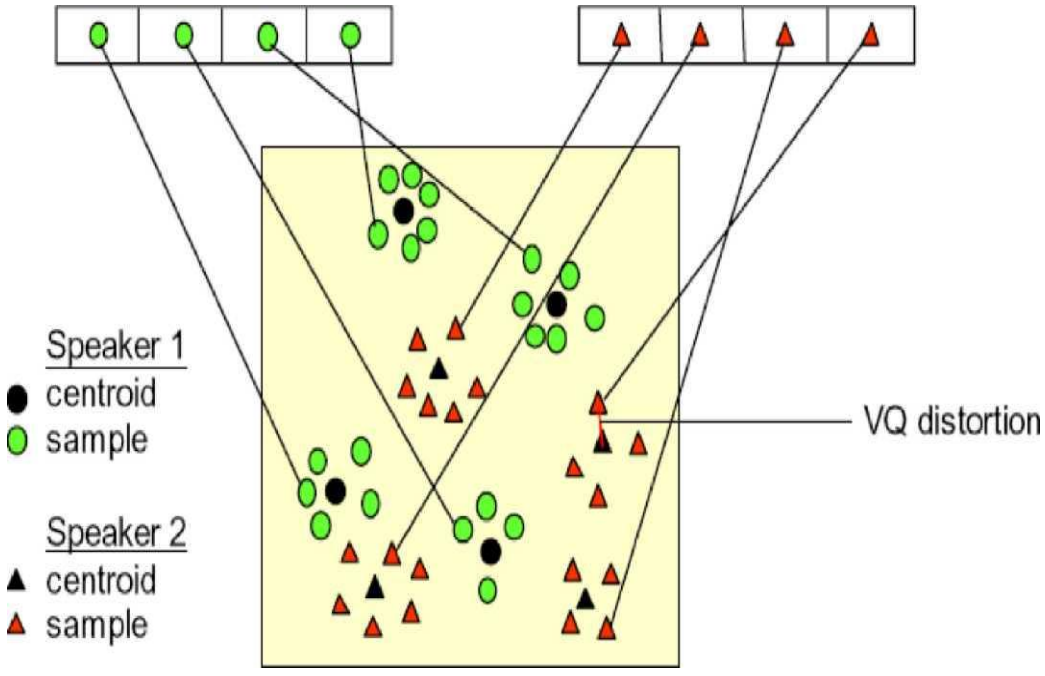

Figure 2.2: Conceptual diagram illustrating vector quantization codebook formation.

One speaker can be discriminated from another based of the location of centroids.

Block Diagram

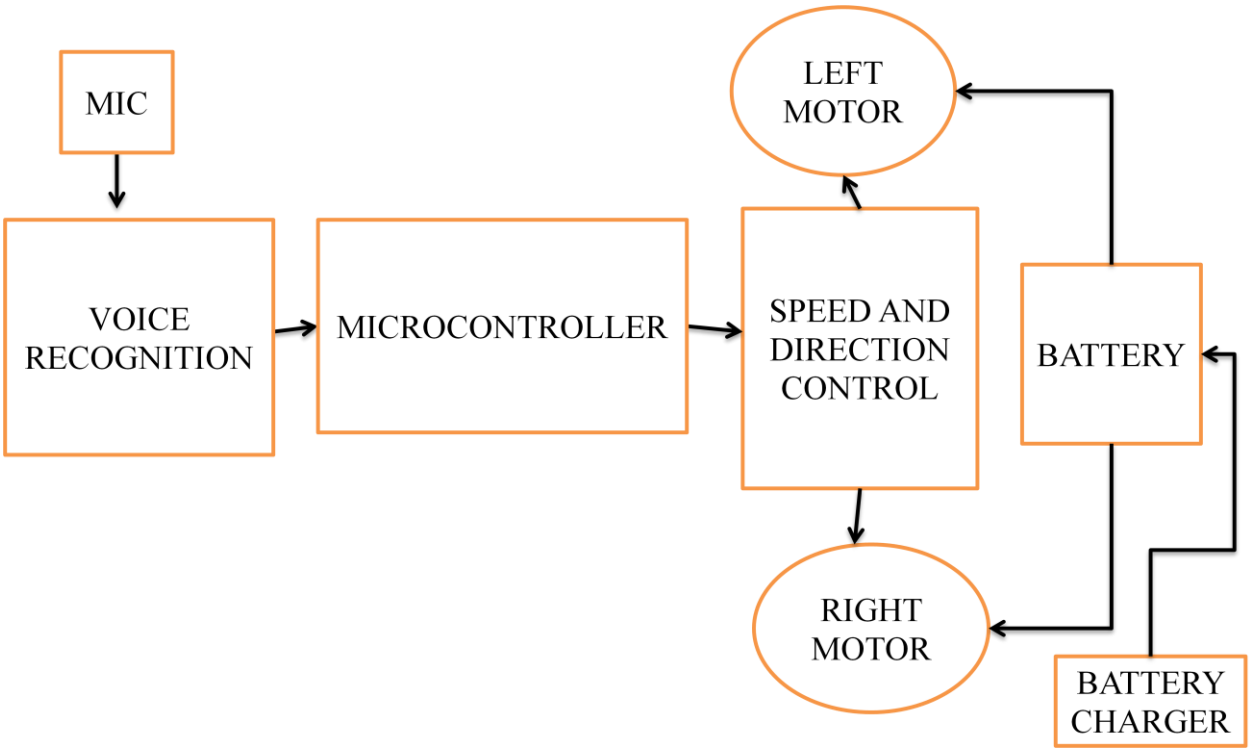

Figure 2.3: Block diagram of the project

Chapter 3

Algorithm Used

This part introduces some definitions and information which will be involved in this project. By concerning and utilizing the theoretic knowledge, we achieved the aim of this project by including DC level and sampling theory, DFT, FFT, spectrum normalization, the cross-correlation algorithm, the autocorrelation algorithm to get the desired signals.

3.1 The DC Level and Sampling Theory

When doing the signal processing analysis, the information of the DC level for the target signal is not that useful except the signal is applied to the real analog circuit, such as AD convertor, which has the requirement of the supplied voltage. When analyzing the signals in frequency domain, the DC level is not that useful. Sometimes the magnitude of the DC level in frequency domain will interfere the analysis when the target signal is most concentrated in the low frequency band.

In WSS condition for the stochastic process, the variance and mean value of the signal will not change as the time changing. So we try to reduce this effect by deducting of the mean value of the recorded signals. This will remove the zero frequency components for the DC level in the frequency spectrum.

In this project, since using the microphone records the person's analog speech signal through the computer, so the data quality of the speech signal will directly decide the quality of the speech recognition. And the sampling frequency is one of the decisive factors for the data quality.

Generally, the analog signal can be represented as

$$x(t) = \sum_{i=1}^{N} A_i \cos(2\pi f_i + \phi_i) \tag{3.1}$$

This analog signal actually consists of a lot of different frequency components. Assuming there is only one frequency component in this analog signal, and it has no phase shift. So this analog signal becomes:

$$x(t) = A\cos(2\pi ft) \tag{3.2}$$

The analog signal cannot be directly applied in the computer. It is necessary to sample the analog signal x(t) into the discrete-time signal x(n), which the computer can use to process.

According to the sampling theorem (Nyquist theorem), when the sampling frequency is larger or equal than 2 times of the maximum of the analog signal frequencies, the discrete-time signal is able to be used to reconstruct the original analog signal. And the higher sampling frequency will result the better sampled signals for analysis. Relatively, it will need faster processor to process the signal and respect with more data spaces.

In non-telecommunications applications, in which the speech recognition subsystem has access to high quality speech, sample frequencies of 10 kHz, 14 kHz and 16 kHz have been used. These sample frequencies give better time and frequency resolution.

In this project, for MATLAB program, the sampling frequency is set as 16 kHz. So the length of the recorded signal in 2 second will be 32000 time units in MATLAB.

3.2 Spectrum Normalization

After doing DFT and FFT calculations, the investigated problems will be changed from the discrete-time signals x(n) to the frequency domain signal $X(\omega)$. The spectrum of the $X(\omega)$ is the whole integral or the summation of the all frequency components. When talking about the speech signal frequency for different words, each word has its frequency band, not just a single frequency and in the frequency band of each word, the spectrum $(|X(\omega)|)$ has its maximum value and minimum value. When comparing the differences between two different speech signals, it is hard or unconvincing to compare two spectrums in different measurement standards. So using the normalization can make the measurement standard the same.

In some sense, the normalization can reduce the error when comparing the spectrums, which is good for the speech recognition. So before analyzing the spectrum differences for different words, the first step is to normalize the spectrum $X(\omega)$ by the linear normalization.

The equation of the linear normalization is as below:

$$y=(x-\text{MinValue})/(\text{MaxValue}-\text{MinValue}) \quad (3.3)$$

After normalization, the values of the spectrum $X(\omega)$ are set into interval [0, 1]. The normalization just changes the values' range of the spectrum, but not changes the shape or the information of the spectrum itself. So the normalization is good for spectrum comparison.

Using MATLAB gives an example to see how the spectrum is changed by the linear normalization.

Firstly, record a speech signal and do the FFT of the speech signal. Then take the absolute values of the FFT spectrum. The FFT spectrum without normalization is as below:

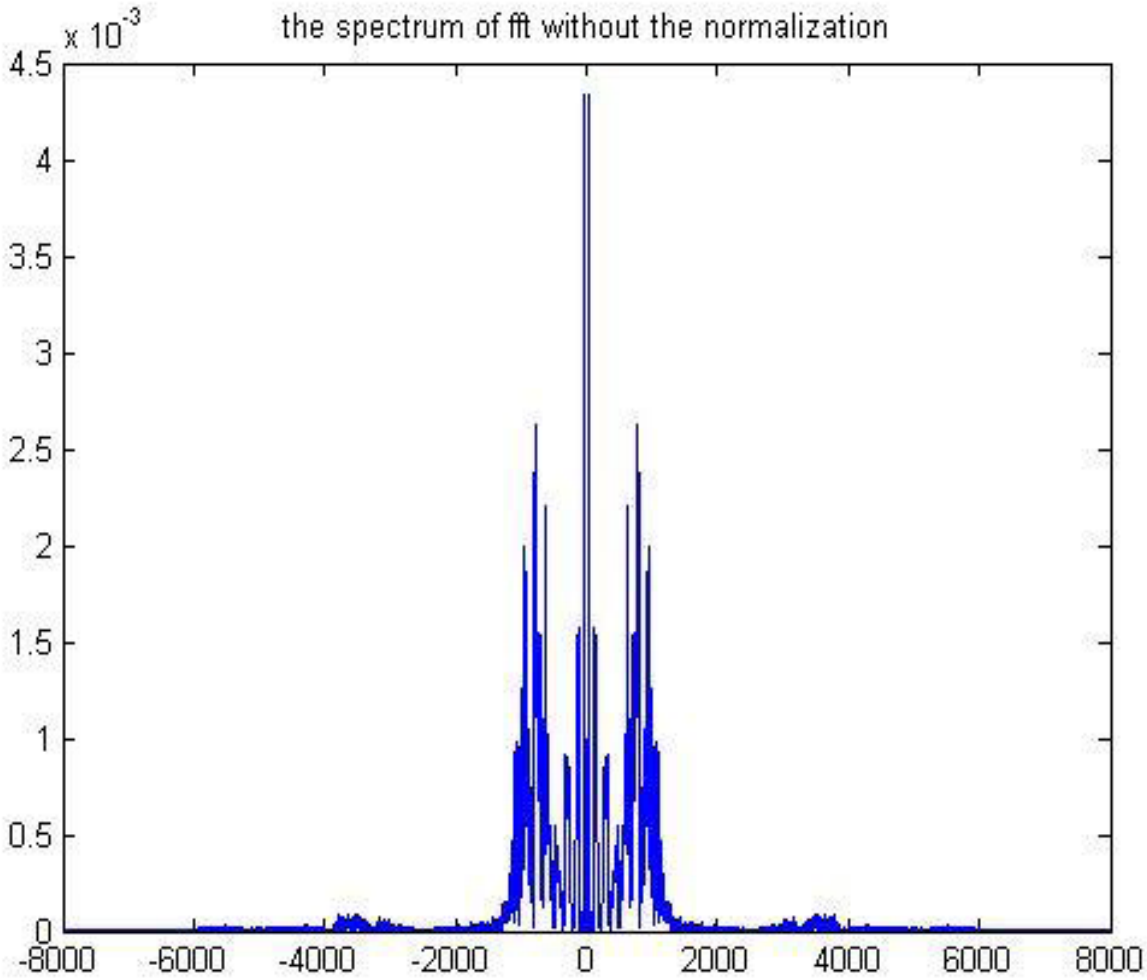

Figure 3.1: Absolute values of the FFT spectrum without normalization

Secondly, normalize the above spectrum by the linear normalization. The normalized spectrum is as below:

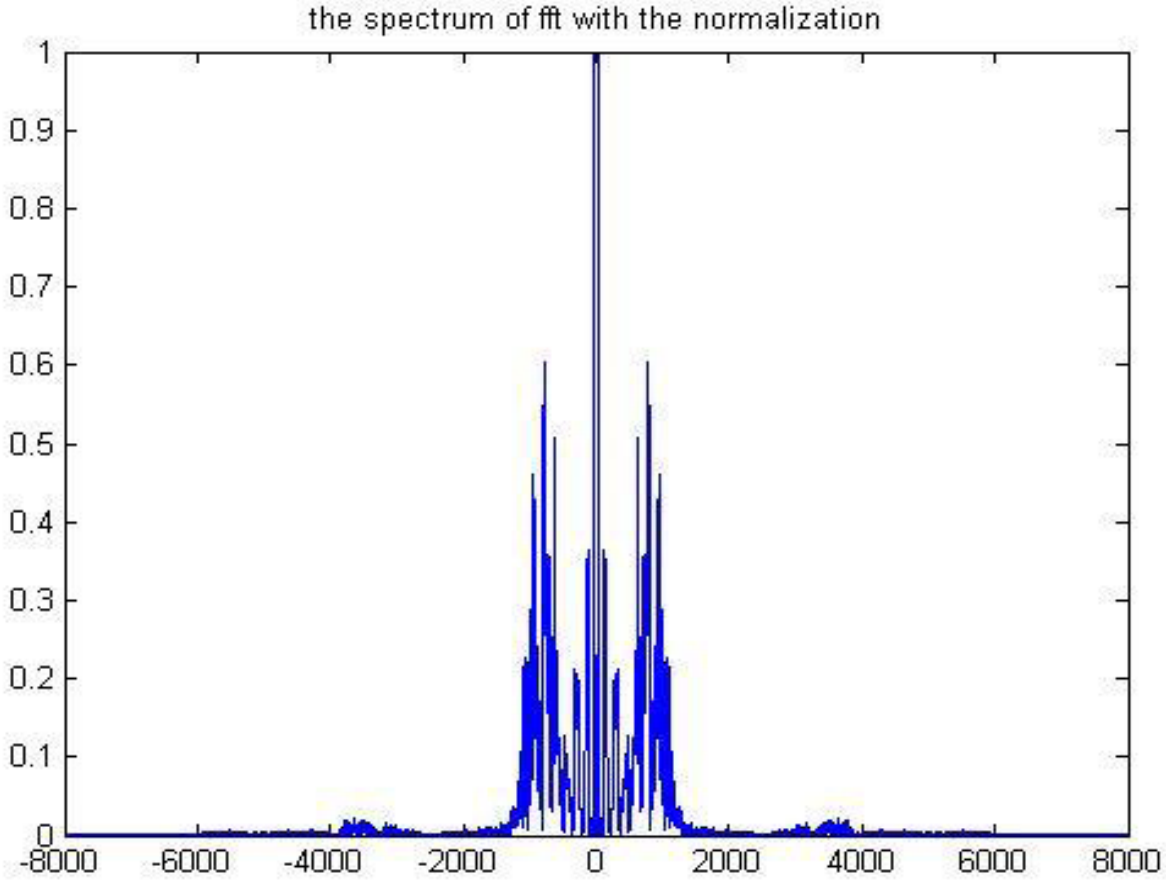

Figure 3.2: Absolute values of the FFT spectrum with normalization

From the Figure.3.1 and the Figure.3.2, the difference between two spectrums is only the interval of the spectrum $X(\omega)$ values, which is changed from [0, 4.5×10-3] to [0, 1]. Other information of the spectrum is not changed.

After the normalization of the absolute values of FFT, the next step of programming the speech recognition is to observe spectrums of the three recorded speech signals and find the algorithms for comparing differences between the third recorded target signal and the first two recorded reference signals.

3.3 The Cross-correlation Algorithm

There is a substantial amount of data on the frequency of the voice fundamental in the speech of speakers who differ in age and sex.

For the same speaker, the different words also have the different frequency bands which are due to the different vibrations of the vocal cord. And the shapes of spectrums are also different.

These are the bases of this project for the speech recognition. In this project, to realize the speech recognition, there is a need to compare spectrums between the third recorded signal and the first two recorded reference signals. By checking which of two recorded reference signals better matches the third recorded signal, the system will give the judgment that which reference word is again recorded at the third time.

When thinking about the correlation of two signals, the first algorithm that will be considered is the cross-correlation of two signals. The cross-correlation function method is really useful to estimate shift parameter. Here the shift parameter will be referred as frequency shift.

The definition equation of the cross-correlation for two signals is as below:

$$r_{xy} = r(m) = \sum_{n=-\infty}^{\infty} x(n)y(n+m) \qquad (3.4)$$

Where $m = 0, \pm1, \pm2, \pm3,....$

From the equation, the main idea of the algorithm for the cross-correlation is approximately 3 steps：

Firstly, fix one of the two signals x(n) and shift the other signal y(n) left or right with some time units.

Secondly, multiply the value of x(n) with the shifted signal y(n+m) position by position.

At last, take the summation of all the multiplication results for x(n) · y(n+m).

For example, two sequence signals x(n) = [0 0 0 1 0], y(n)= [0 1 0 0 0], the lengths for both signals are N=5. So the cross-correlation for x(n) and y(n) is as the following figures shown:

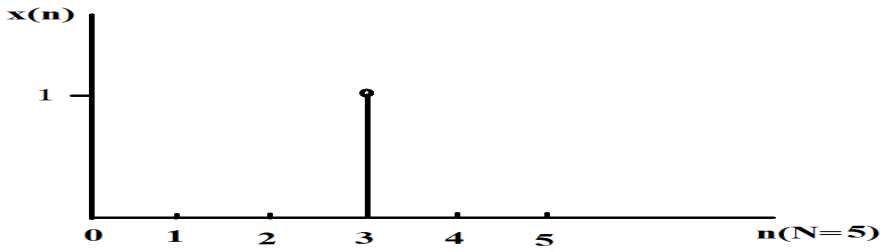

Figure 3.3: The signal sequence x(n)

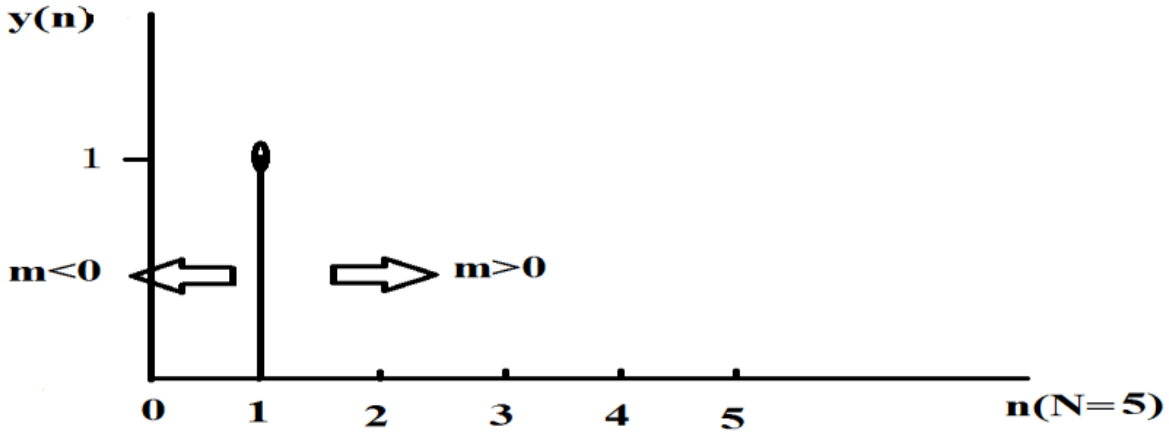

Figure 3.4: The signal sequence y(n) will shift left or right with m units

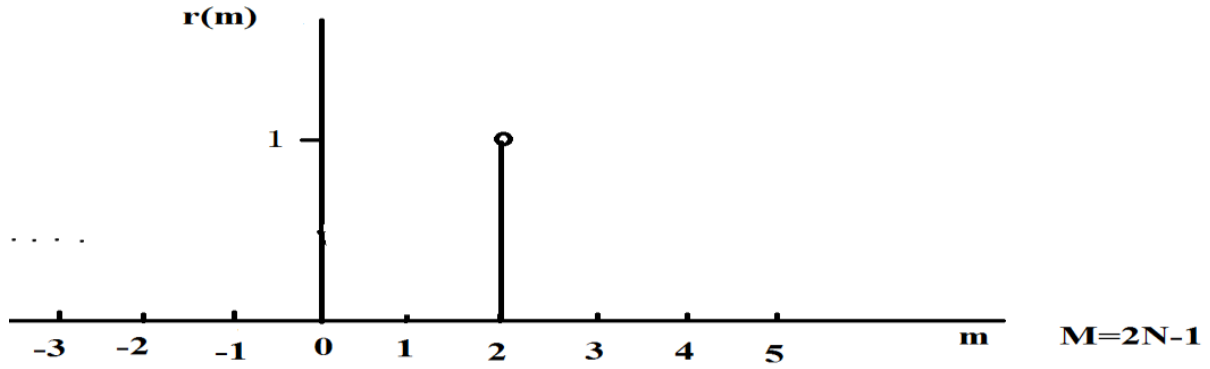

Figure 3.5: The results of the cross-correlation, summation of multiplications

As the example given, there is a discrete time shift about 2 time units between the signals x(n) and y(n). From Fig 4.5, the cross-correlation r(m) has a non-zero result value, which is equal 1 at the position m=2. So the m-axis is no longer the time axis for the signal. It is the time-shift axis. Since the lengths of two signals x(n) and y(n) are both N=5, so the length of the time-shift axis is 2N.

When using MATLAB to do the cross-correlation, the length of the cross-correlation is still 2N. But in MATLAB, the plotting of the cross-correlation is from 0 to 2N-1, not from –N to +N anymore. Then the 0 time-shift point position will be shifted from 0 to N.

So when two signals have no time shift, the maximum value of their cross-correlation will be at the position m=N in MATLAB, which is the middle point position for the total length of the cross-correlation.

From the example, two important information of the cross-correlation can be given. One is when two original signals have no time shift, their cross-correlation should be the maximum; the other information is that the position difference between the maximum value position and the middle point position of the cross-correlation is the length of time shift for two original signals.

Now assuming the two recorded speech signals for the same word are totally the same, so the spectrums of two recorded speech signals are also totally the same. Then when doing the cross-correlation of the two same spectrums and plotting the cross-correlation, the graph of the cross-correlation should be totally symmetric according to the algorithm of the cross-correlation.

However, for the actual speech recording, the spectrums of twice recorded signals which are recorded for the same word cannot be the totally same. But their spectrums should be similar, which means their cross-correlation graph should be approximately symmetric. This is the most important concept in this project for the speech recognition.

By comparing the level of symmetric property for the cross-correlation, the system can make the decision that which two recorded signals have more similar spectrums. In other words, these two recorded signals are more possibly recorded for the same word.

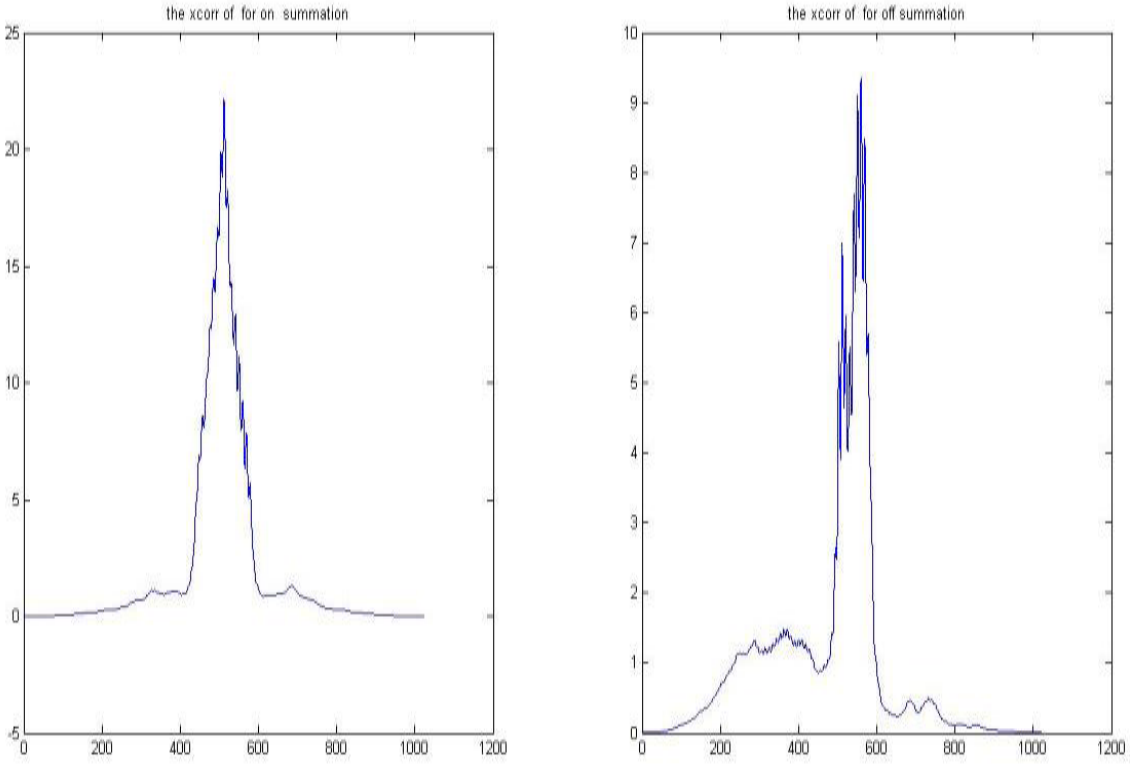

Figure 3.6: The graphs of the cross-correlations

The first two recorded reference speech words are “hahaha” and “meat”, and the third time recorded speech word is “hahaha” again. From Figure 3.6, the first plotting is about the cross-correlation between the third recorded speech signal and the reference signal “hahaha”. The second plotting is about the cross-correlation between the third recorded speech signal and the reference signal “meat”. Since the third recorded speech word is “hahaha”, so the first plotting is really more symmetric and smoother than the second plotting.

For the speech recognition comparison, after calculating the cross-correlation of two recorded frequency spectrums, there is a need to find the position of the maximum value of the cross-correlation and use the values right to the maximum value position to minus the values left to the maximum value position. Take the absolute value of this difference and find the mean square-error of this absolute value. If two signals better match, then the cross-correlation is more symmetric. And if the cross-correlation is more symmetric, then the mean square-error should be smaller. By comparing of this error, the system decides which reference word is recorded at the third time.

3.4 The Auto-correlation Algorithm

In the previous part, it is about the cross-correlation algorithm. See the equation , the autocorrelation can be treated as computing the cross-correlation for the signal and itself instead of two different signals. This is the definition of auto-correlation in MATLAB. The auto-correlation is the algorithm to measure how the signal is self-correlated with itself.

The equation for the auto-correlation is:

$$r_x(k) = r_{xx}(k) = \sum_{k=-\infty}^{\infty} x(n)x(n+k) \tag{3.5}$$

The figure below is the graph of plotting the autocorrelation of the frequency spectrum $X(\omega)$

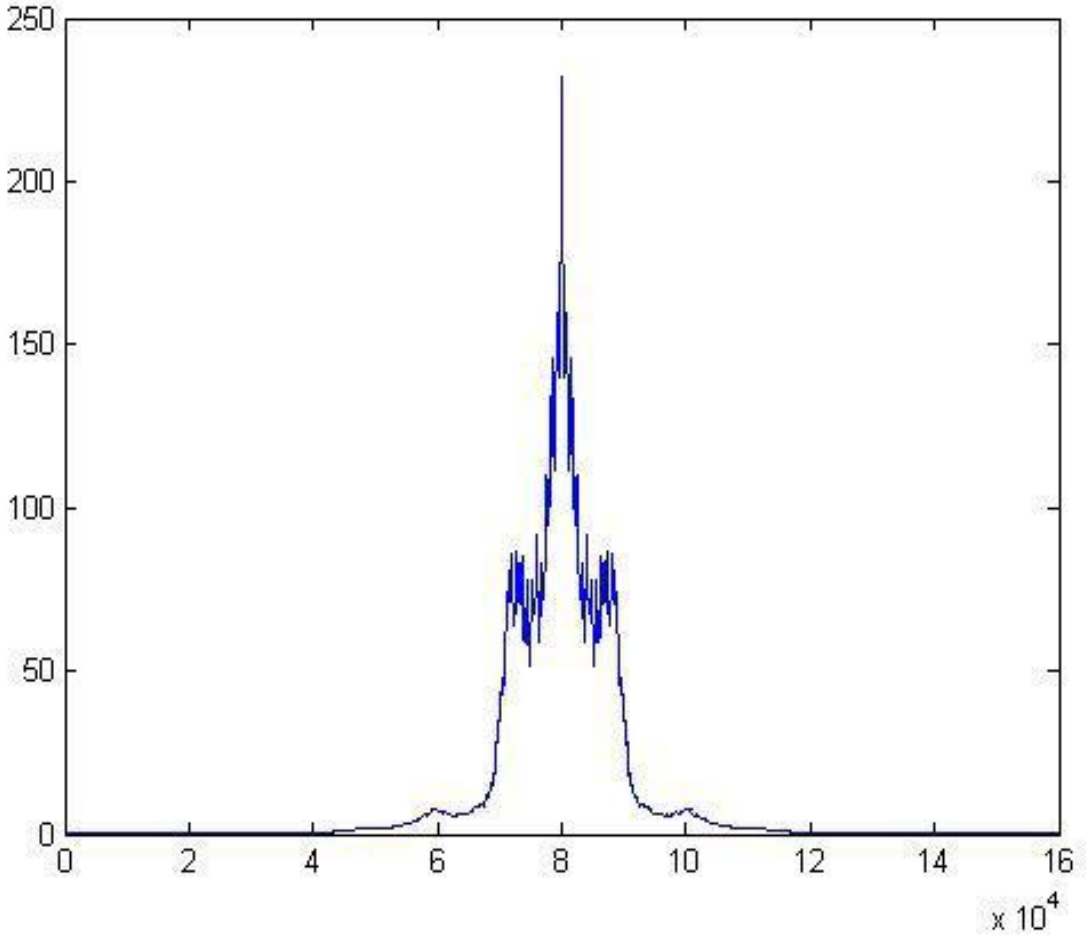

Figure 3.7 The autocorrelation for $X(\omega)$

3.5 Use of spectrogram Function in MATLAB to Get Desired Signals

The spectrogram is a time-frequency plotting which contains power density distribution at the same time with respect to both frequency axis and time axis.

In MATLAB, it is easy to get the spectrogram of the voice signal by defining some variables: the sampling frequency, the length of Shot-Time Fourier Transform (STFT) and the length of window.

The STFT is firstly to use the window function to truncate the signal in the time domain, which makes the time-axis into several parts.

If the window is a vector, then the number of parts is equal to the length of the window. Then compute the Fourier Transform of the truncated sequence with defined FFT length (nfft).

Chapter 4

Simulations and Results

4.1 Programming Steps

(1) Initialize the variables and set the sampling frequency fs=16000.
Use "wavrecord" command to record 3 voice signals. Make the first two recordings as the reference signals. Make the third voice recording as the target voice signal.

(2) Use "spectrogram" function to process recorded signals and get returned matrix signals.

(3) Transpose the matrix signals for rows and columns, take "sum" operation of the matrix and get a returned row vector for each column summation result. This row vector is the frequency spectrum signal.

(4) Normalize the frequency spectrums by the linear normalization.

(5) Do the cross-correlations for the third recorded signal with the first two recorded reference signals separately.

(6) This step is important since the comparison algorithm is programed here. Firstly, check the frequency shift of the cross-correlations. Here it has to be announced that the frequency shift is not the real frequency shift. It is processed frequency in MATLAB. By the definition of the spectrum for the "nfft", which is the length of the STFT programmed in MATLAB, the function will return a frequency range which is respect to the "nfft". If

"nfft" is odd, so the returned matrix has $\frac{nfft+1}{2}$ rows; if "nfft" is even, then the returned matrix has $\frac{nfft}{2}+1$ rows. These are defined in MATLAB. Rows of the returned "spectrogram" matrix are still the frequency ranges.

If the difference between the absolute values of frequency shifts for the two cross-correlations is larger or equal than 2, then the system will give the judgment only by the frequency shift. The smaller frequency shift means the better match. If the difference between the absolute values of frequency shifts is smaller than 2, then the frequency shift difference is useless according to the experience by large amounts tests. The system needs continuously do the

comparison by the symmetric property for the cross-correlations of the matched signals. The algorithm. According to the symmetric property, MATLAB will give the judgment.

4.2 Simulations

This is the MATLAB [8] command window that we see after the initialization of our ASR system Programme.

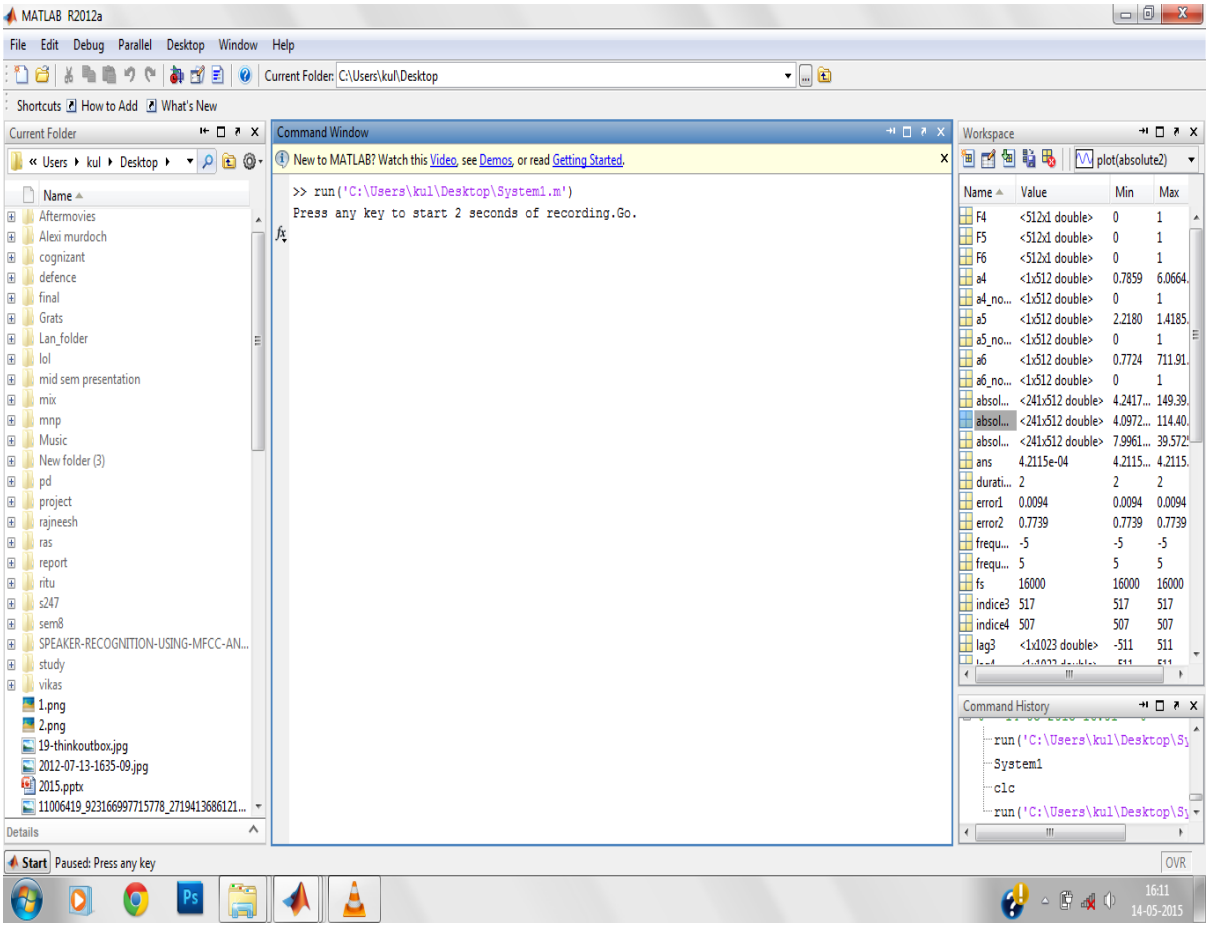

Figure 4.1: Recording of Go Command

In the Command window user has to say GO through Mic, after pressing any key on the keyboard in 2 second. This will create a Database of the sound. A wav file will simultaneously create in the folder in which the programme is running. The wav file is created by the name of move.wav.

Similarly, after the entry of Go command in the database, the user has to create one more entry of STOP command in the data base, in the similar manner as done for the go command.

The wav file is created by the name of stop.wav.

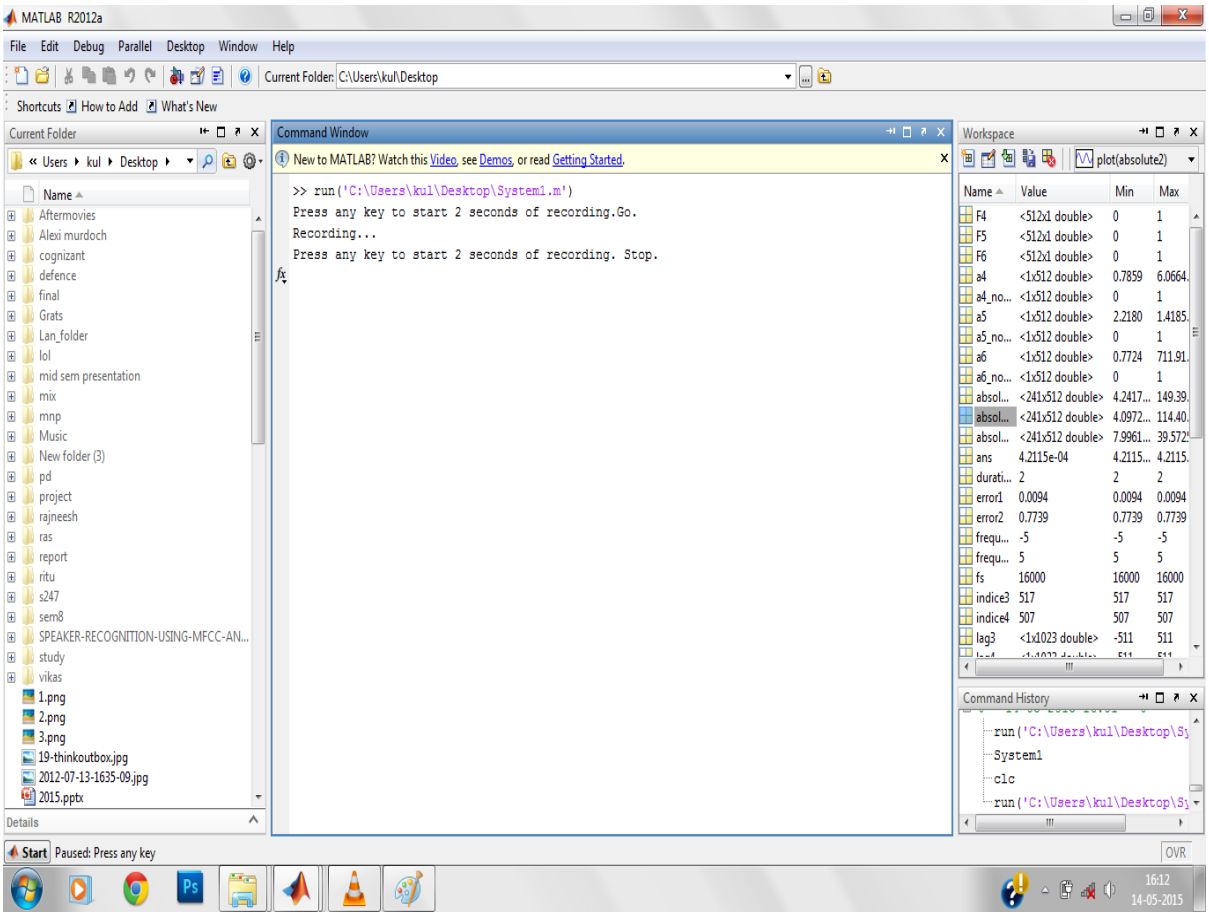

Figure 4.2: Recording of Stop Command

The next step is the User command, mean the recording of the command given by the user weather the user want the move or stop. The recording of the user's command is done in the same way as done in the earlier processes. The user command, weather it is Go or Stop, will match itself from the earlier entered data in the data base using the Cross- Correlation algorithms and the results will be displayed according to the match.

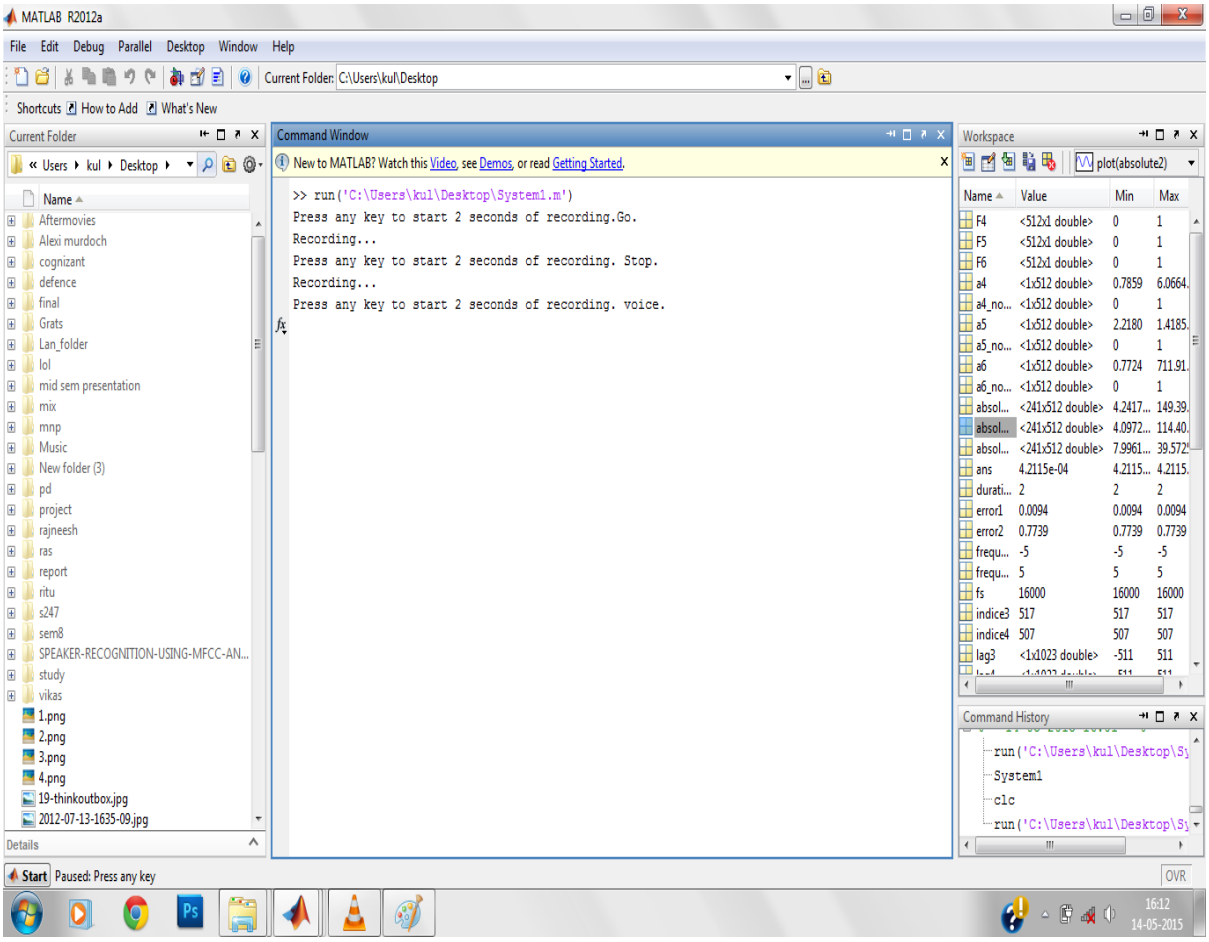

Figure 4.3: Recording of User's Command

4.3 Results

The information of the first statistical simulation results for our ASR system [y] are displayed after the entry of the database and the matching process of the user's command. The most important thing to know is that the environment in which the user has recorded and commanded must be silent. Otherwise there will be addition of noise in the recording and this will hinder the matching process and even the result will get change.

The Figure 4.4 is about frequency spectrums for three recorded signals, but the axis is not the real frequency axis since the figure is got by STFT.

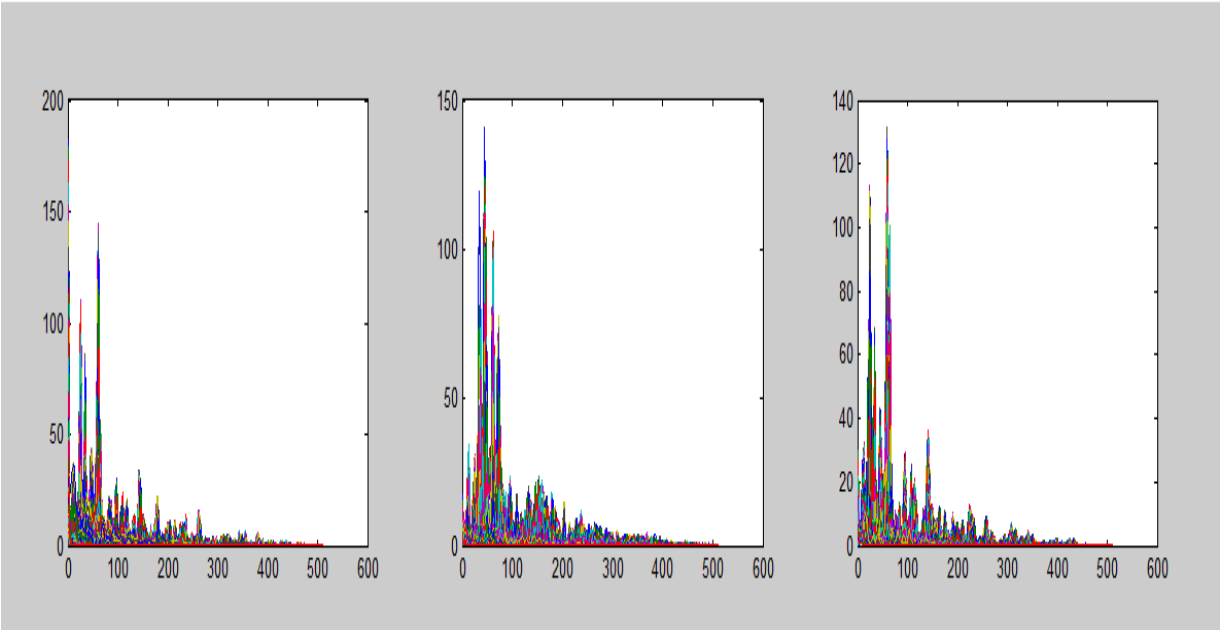

Figure 4.4: Frequency Spectrum of the GO STOP and User's Command without Normalisation.

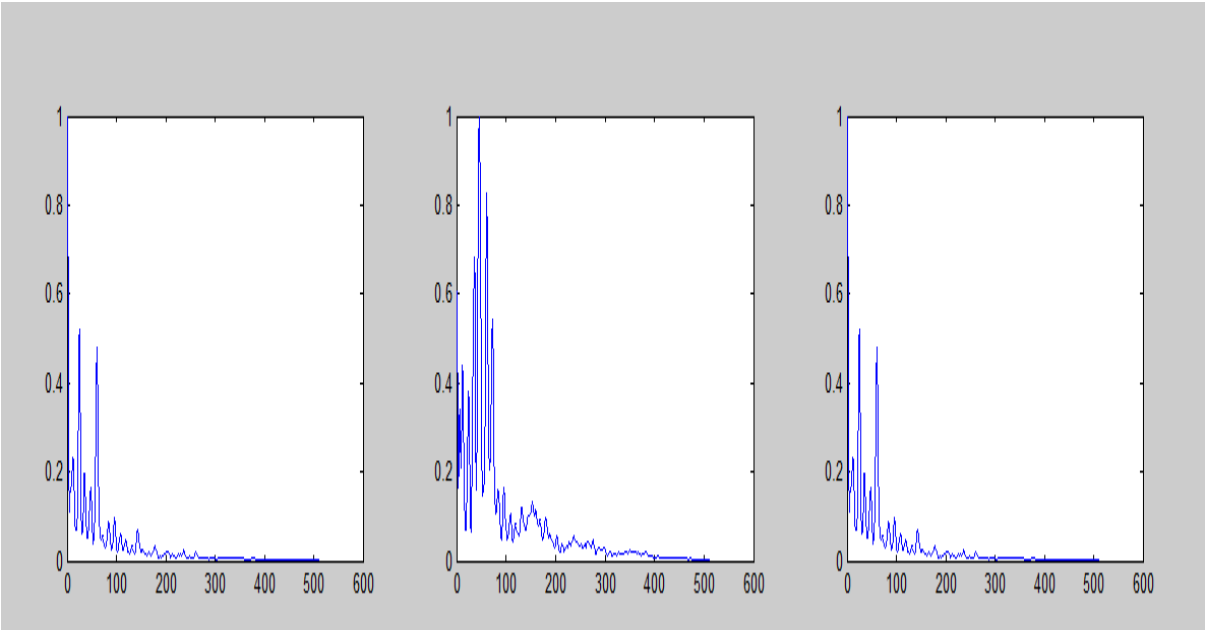

Figure 4.5: Frequency Spectrum of the GO STOP and User's Command after Normalisation.

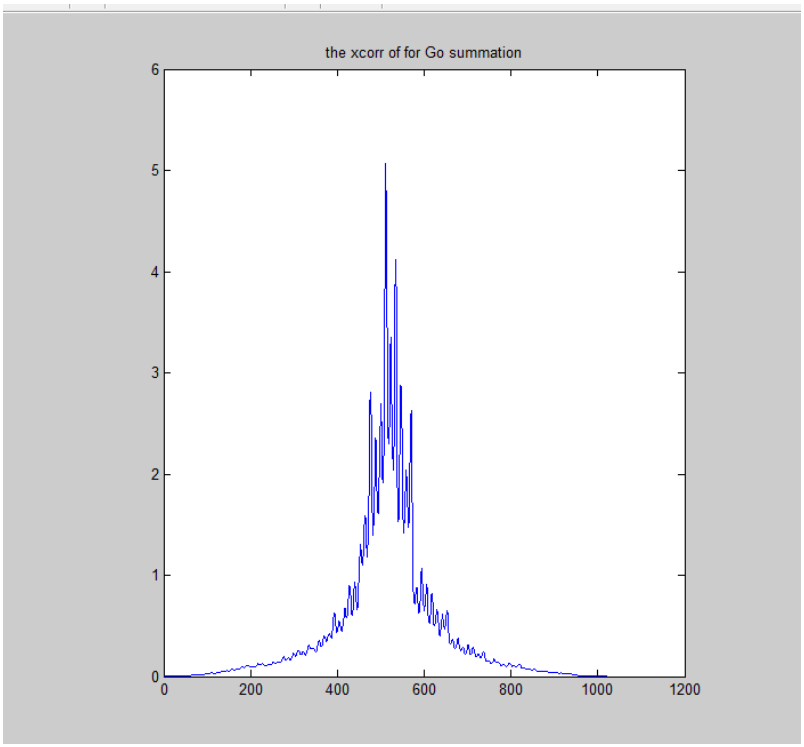

Figure 4.6: The Cross-Correlation of the GO command with the User's command

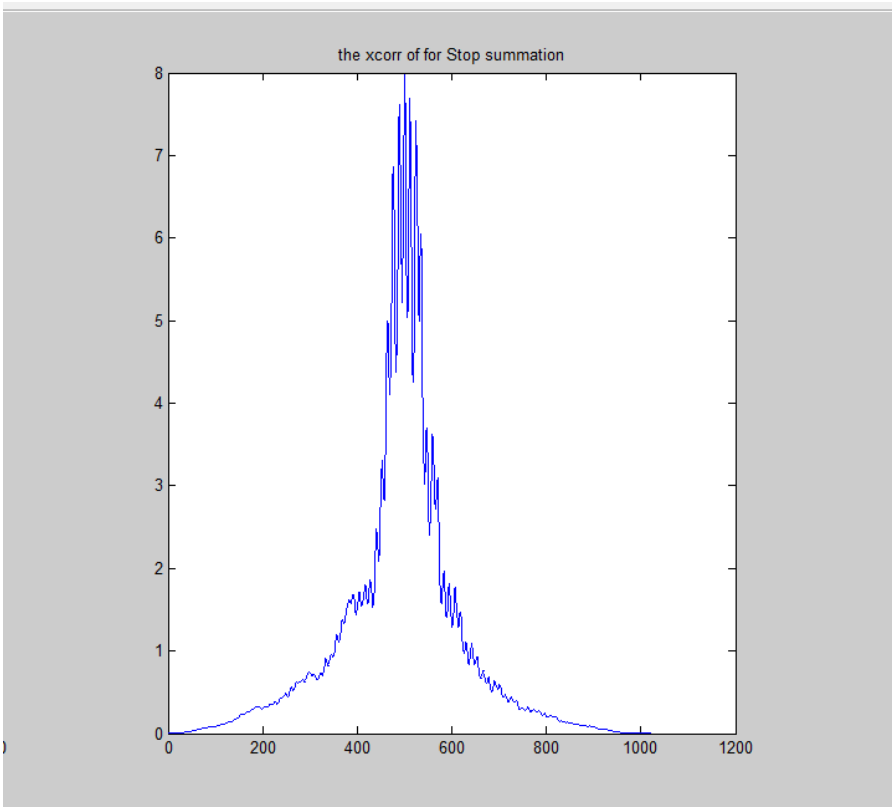

Figure 4.7: The Cross-Correlation of the STOP command with the User's command

There are two method used for the Signal matching in the cross-correlation algorithm

- The frequency shift method.
- The symmetric property of cross-correlation Graph for the matched signal shape.

Firstly the check is about the frequency shift. The judgement is only good when frequency shift difference is larger or equal to 2. The smaller of the frequency shift, the better match signals. The frequency shift is not reliable for the words having really close pronunciations, like "on" and "off" are really pronounced close.

At this situation, the designed system will give the judgments by comparing the errors of the symmetric property of the cross-correlations.

Chapter 5

Conclusion and Future Scope

5.1 Conclusion

The goal of our project was to create a speaker recognition system, and apply it to a set of instructions given by the person sitting on the wheelchair. By investigating the extracted features of the unknown speech and then compare them to the stored extracted features for each different speaker in order to identify the instruction.

The feature extraction is done by using cross-correlation Technique. In the recognition stage, we used the frequency as well as the symmetric property of cross-correlation for the matching of signal shape.

During this project, we have found out that the cross-correlation Technique provides us with the faster speaker identification process than any other algorithms such as MFCC approach and FFT approach.

5.2 Comparison

We successfully simulated the Cross-correlation Technique approach using MATLAB. We concluded from our experiments that for an isolated word recognition system like the one we aimed to implement in our project, the Cross-correlation Technique approach proved to be more effective. The reason for this is that, the MFCC approach has a drawback. As explained earlier the key to this approach is using the energies, however, this may not be the best approach as was discovered.

It was seen from the experiments that because of the prominence given to energy, this approach failed to recognize the same word uttered with different energy. Also, as this takes the summation of the energy within each triangular window it would essentially give the same value of energy irrespective of whether the spectrum peaks at one particular frequency and falls to lower values around it or whether it has an equal spread within the window.

However, as the time domain approach is not entirely based on energy but also based on the dominant frequencies within small segments of speech, it makes good use of the quasi - stationary property of speech. But the results given by the cross-correlation Technique approach is more suitable and reliable for the isolated work recognizer required for a voice based biometric system.

5.3 Other Applications of our ASR system

After nearly sixty years of research, speech recognition technology has reached a relatively high level. However, most state-of-the-art ASR systems run on desktop with powerful microprocessors, ample memory and an ever-present power supply. In these years, with the rapid evolvement of hardware and software technologies, ASR has become more and more expedient as an alternative human-to-machine interface that is needed for the following application areas:

- Stand-alone consumer devices such as wrist watch, toys and hands-free mobile phone in car where people are unable to use other interfaces or big input platforms like keyboards are not available.
- Single purpose command and control system such as voice dialing for cellular, home, and office phones where multi-function computers (PCs) are redundant.

Some of the applications of speaker verification systems are:

- Time and Attendance Systems
- Access Control Systems
- Telephone-Banking/Broking
- Biometric Login to telephone aided shopping systems
- Information and Reservation Services
- Security control for confidential information
- Forensic purposes

Voice based Telephone [9] dialing is one of the applications we simulated. The key focus of this application is to aid the physically challenged in executing a mundane task like telephone dialing. Here the user initially trains the system by uttering the digits from 0 to 9. Once the system has been trained, the system can recognize the digits uttered by the user who trained the system. This system can also add some inherent security as the system based on Cross- Correlation approach is speaker' word dependent.

Presently systems have also been designed which incorporate Speech and Speaker Recognition. Typically a user has two levels of check. He/She has to initially speak the right password to gain access to a system.

5.4 Scope for future work

This project focused on "Isolated Word Recognition". But we feel the idea can be extended to "Continuous Word Recognition"[10] and ultimately create a Language Independent Recognition System based on algorithms which make these systems robust. The use of Statistical Models like HMMs, GMMs or learning models like Neural Networks and other associated aspects of Artificial Intelligence can also be incorporated in this direction to improve upon the present project. This would make the system much tolerant to variations like accent and extraneous conditions like noise and associated residues and hence make it less error prone.

Here we used cross-correlation Technique. But someone can try k-means algorithm also. This field is very vast and research is also done for that purpose.

Referrences

[1] Lawrence Rabiner, Biing-Hwang Juang – „*Fundamentals of Speech Recognition'*

[2] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun – „*An Efficient MFCC Extraction Method in Speech Recognition'*, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong, IEEE – ISCAS, 2006

[3] Waleed H. Abdulla – „*Auditory Based Feature Vectors for Speech Recognition*

Systems', Electrical & Electronic Engineering Department, The University of Auckland

[4] Beth Logan – „*Mel Frequency Cepstral Coefficients for Music Modeling'*, Cambridge Research Laboratory, Compaq Computer Corporation

[5] Woszczyna, M.: "JANUS 93: Towards Spontaneous Speech Translation", IEEE Electronics & communication Eng. Institute of technology, Nirma University _ Page 67 Proceedings Conference on Neural Networks, (1994).

[6] Nilsson, M.; Ejnarsson, M.: "Speech Recognition Using HMM: Performance

Evaluation in Noisy Environments", MS Thesis, Blekinge Institute of Technology,

Department of Telecommunications and Signal Processing, (2002).

[7] www.dspguide.com/zipped.htm: "The Scientist and Engineer's Guide to Digital

Signal Processing" (Access date: March 2005).

[8] Brookes, M.: "VOICEBOX: a MATLAB toolbox for speech processing",

www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, (2003).

[9] Skowronski, M.D.: "Biologically Inspired Noise-Robust Speech Recognition for

Both Man and Machine", PhD Thesis, The Graduate School of the University of

Florida, (2004).

[10] Davis, S.; Mermelstein, P.: “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 4 (1980).